**Ludwig-Maximilians-Universität München**
**Institut für Informatik**
Dr. Eirini Ntoutsi
PD Dr. Matthias Schubert

# Knowledge Discovery in Databases II

WS 2015/2016

## Übungsblatt 9: Ensembles

### Aufgabe 9-1 Variance Computation

Given the following 1-dimensional data set:

$$A = \{0, +1, -1, +2, -2, +3, -3, \ldots, +100, -100\} \quad B = \{a_i + 10^{10}\} \quad C = \{a_i \cdot 10^{-10} + 1\}$$

Compute the variance of the data using:

- The naive method using 2 iterations: compute the mean first and then the squared average difference between the data and the mean

$$\mu := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{Var}(X) := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Compute the variance in one iteration based on the sum and square sum:

$$\text{Var}(X) := \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

- Knuth and Welfort' algorithm:

```
n = 0
mean = 0.0
sum2 = 0.0
for x in data:
        n = n + 1
        delta = x - mean
        mean = mean + delta/n
        sum2 = sum2 + delta*(x - mean)
var = sum2 / (n-1)
```

- Run the numpy code.
- Show that each method is mathematically equivalent.
  (Tipps: Cancel the term $\frac{1}{n-1}$ as soon as possible. For the third method use complete induction and determine suitable invariants.)

Use a precision of `double` for your computation.
What can be observed? How can the observed phenomenon be explained.

## Aufgabe 9-2 Ensemble Multi-Class-Classification

We have previously considered the ensemble strategies *one-versus-rest*, *all-pairs*, and *ECOC*. These have allowed us to reduce multi-class classification problems to multiple two-class classification problems. For *one-versus-rest* und *all-pairs* the application/test step was a simple majority voting, *ECOC* required a more sophisticated decision rule.

A further approach is given by the DDAG-strateggy: Individual *all-pairs*-classifiers form a directed, acyclic graph (DDAG=*Decision Directed Acyclic Graph*) to facilitate the classification result. See the following figure:

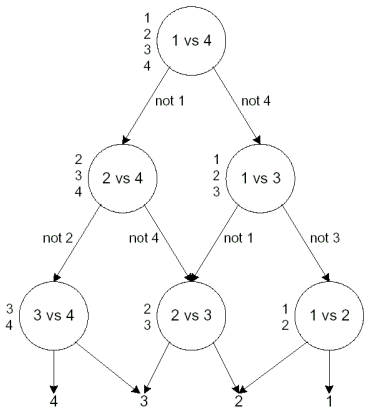


Abbildung 1: Classification strategy DDAG

(a) What advantages and disadvantages does this strategy have compared to voting using pairwise classifiers?

(b) For each base-classifier, assume a complexity given by a function $t : \mathbb{N} \to \mathbb{R}_0^+$, which is dependent on the number of training samples. How do different strategies perform regarding the time requirements in the training phase for $n$ classes and $m$ samples in each class? How do they perform in the application phase, assuming constant time for a prediction of an single base-classifier?

## Aufgabe 9-3 Error Correcting Output Codes

In this exercise, we will implement an ensemble classifiers for the sklearn library which is based on a GaussianNB classifier and a error correcting output codes.

(a) Download and inspect the template ecoc_template.py. I contains a template for implementing a new classifier

(b) Implement the new ensemble classifier.

(c) Test the new classifier on the provided arff file or different numbers of classifiers.